Legal Language and Legal Knowledge Management Applications

Giulia Venturi

Institute of Computational Linguistics, CNR
via G. Moruzzi 1, 56124, Pisa, Italy
giulia.venturi@ilc.cnr.it
http://www.ilc.cnr.it

Abstract. This work is an investigation into the peculiarities of legal language with respect to ordinary language. Based on the idea that a shallow parsing approach can help to provide enough detailed linguistic information, this work presents the results obtained by shallow parsing (i.e. chunking) corpora of Italian and English legal texts and comparing them with corpora of ordinary language. In particular, this paper puts the emphasis of how understanding the syntactic and lexical characteristics of this specialised language has practical importance in the development of domain–specific Knowledge Management applications.

Keywords: Parsing Legal Texts, Natural Language Processing, Legal Language, Knowledge Management Applications.

1 Introduction

This work is an investigation into the peculiarities of legal language with respect to ordinary language. Within the specialised–domain field, the increasing need for text processing of large document collections has prompted research efforts devoted to automatically processing the sublanguage used in domain-specific corpora. Interestingly, sublanguage processing encompasses not only the numerous inherent complexities of ordinary Natural Language Processing (NLP), but also the treatment of domain–specific peculiarities. Accordingly, beyond the general NLP difficulties, the specificities of domain–specific features make the automatic processing of these kind of corpora a challenging task and demand specific solutions. This is the reason why this article puts the emphasis of how understanding the characteristics of a specialised language has practical importance in the development of domain–specific Knowledge Management applications.

More than within other research communities, the reciprocal exchange between the bio–medical community and the Natural Language Processing one is a promising example in this direction. The very active research field (see [3] for an updated overview) witnesses that this joint effort between the two is fruitful for the purposes of both communities. It follows that on the one hand a variety of NLP techniques are exploited for a number of domain–specific applications such as Biological Text Mining, the construction of ontological resources, Information Retrieval, and Event Extraction. On the other hand, various efforts are

E. Francesconi et al. (Eds.): Semantic Processing of Legal Texts, LNAI 6036, pp. 3–26, 2010.

also devoted to investigating biomedical language characteristics with a view to customise NLP tools, in order to adapt them to the processing of biomedical sub–language.

On the contrary, within the Artificial Intelligence and Law community (AI& Law), to the author's knowledge, little attention has been devoted both to techniques coming from the NLP community, and to research efforts concerned with the analysis of legal language peculiarities. In particular, it has been overlooked how understanding the characteristics of this specialised language can help to shed light on the main difficulties of extracting *semantic information* out of legal documents.

Within the legal domain the situation is made more challenging, with respect to other specialised domains, by the fact that laws are invariably conveyed through natural language. According to linguistic studies, such as Garavelli [20], legal language, although still different from ordinary language, is in fact not dramatically independent from every day speech. It rather makes a *specific use* of lexical and syntactic peculiarities typical of ordinary language. Consequently, it can be seen both as an *extension* and as a *reduction* of the possibilities offered by ordinary language. Moreover, since the 80s the close intertwining between legal knowledge and legal language has been acknowledged within the Artificial Intelligence and Legal Reasoning community. Pointing out some special characteristics of the legal domain, Dyer in Rissland [26] claims that "[m]odeling what a lawyer does is more complex than modeling experts in technical/scientific domains. First, all of these complex conceptualizations are expressed in natural language, so modeling the comprehension ability of the lawyer requires solving the natural language problem". However, unfortunately, as it will be discussed in Section 2, few research activities have focussed on this topic.

According to these premises, this article aims at suggesting how fruitful an analysis of the main legal language peculiarities can be. For this purpose, a comparative study of the specialised language used within legal documents with respect to the newswire language is carried out at a considerable level of detail. In particular, corpora of Italian and English legal texts have been parsed and compared with corpora of ordinary language in order to detect the main syntactic and lexical characteristics of legal language. The eventual goal is to suggest that a study phase of linguistic peculiarities of legal texts can have practical importance in many Legal Knowledge Management tasks.

2 Background and Motivation

Despite the urgent need for legal text semantic processing, according to McCarty [16], little attention has been paid within the AI&Law community to how NLP can contribute to Legal Knowledge Management. Rather, research in the field has been conducted mainly from a top–down perspective, uniquely stemming from domain–theoretical assumptions. So that, a bottom–up investigation of if and to what extent legal text semantic processing may benefit from linguistic analyses and techniques has been mostly overlooked.

Moreover, a corpus of law texts *linguistically annotated* (i.e. with morphological, syntactic and semantic information made explicit) is lacking, even though it would be useful for a number of Knowledge Management applications, such as Information Extraction and Domain Ontology Learning (as it has been recently raised in McCarty [17]).

However, in the last few years the number of NLP–oriented research activities has increased as witnessed by the workshops and tutorials recently organized on this topic. As a matter of fact, a survey of the main Knowledge Management applications in the legal domain can show that NLP tools are currently exploited in various studies. This is the case of the following cases:

1. **Legal Ontology Learning**, carried out by Van Gog et al. in [32], Lame in [11], Saias et al. in [28], Walter et al. in [34] and Völker et al. in [31];
2. **Legal Information Extraction**, carried out by Walter in [35] and McCarty in [16];
3. **Legal Semantic Annotation**, carried out by Bartolini et al. in [5], Brighi et al. in [6] and Spinosa et al. in [29];
4. **Automatic identification of legal terms** for lexicographic purposes, carried out by Pala et al. in [23] and [24];
5. **Legal Knowledge Modeling**, carried out by Nakamura et al. in [21];
6. **Legal Argumentation**, carried out by Moens et al. in [19], by Mochales et al. in [18] and by Wyner et al. in [36]; more recently, Wyner et al. in [37] put the focus on the need for bridging Computational Linguistics and Legal Argumentation efforts. In fact, the use of NLP tools is meant to support the formal construction of argument schemes;
7. **Legal Automatic Summarisation**, carried out by Grover et al. in [9].

However, in these studies and projects little attention has been paid:

1. to take into account the potentialities offered by each level of linguistic analysis (i.e. sentence splitter, tokenization in single word units, morphological analysis and shallow or deep syntactic parsing) to the following semantic processing of legal texts;
2. to put the emphasis on the need for domain–specific customizations asked for legal language peculiarities;
3. to point out legal language peculiarities with a special view to those characteristics which make this specialised language different with respect to ordinary language.

Interestingly enough, the vast majority of these works takes into account only the output of the component in charge of the *syntactic parsing* of the text. Mostly, the research activities considered here take into account the output of the *deep* level of syntactic parsing they have carried out. It follows that the previous levels of linguistic analysis, i.e. sentence splitter, tokenization in single word units and morphological analysis, are overlooked. However, according to what has been noted for the Biomedical Language Processing field in Ananiadou et al. [3], each level of linguistic analysis has been typically associated with specific processing

components aimed at tackling domain–specific phenomena at some level. By focussing on the relationship between different processing components and various kinds of analyses, the authors allow appreciation of how each particular type of component relates to the overall Text Mining task.

In the legal domain, one exception is the analysis carried out by Pala and colleagues in [23], where results of the morphological analysis and lemmatizations of the Penal Code of the Czech Republic are presented. For this purpose, a morphological analyser designed for general Czech has been customised according to some legal language peculiarities, namely by adding legal terms. Interestingly, the authors put the focus on the main outcome of their work: as a result, they have obtained basic knowledge about the grammatical structure of legal texts (law terminology). Starting from the analysis of this processing component (i.e. the morphological analyser), they were further concerned in [24] with the development of a database containing valency frames of legal verbs, suitable for the description of the meanings of legally relevant verbs. In this respect, the Pala and colleagues' effort is aimed at exploring how even the morphological level of linguistic analysis can help the investigation of the semantic nature of legal language.

Secondly, just few of the works aforementioned are explicitly focused on the need for domain–specific customizations needed for legal language peculiarities. As witnessed by the efforts carried out in the Biomedical Language Processing community (see e.g. Lease et al. [12], Pyysalo et al. [25] and Sagae et al. [27]), studies overtly devoted to the adaptation of NLP tools for domain–specific purposes may improve the document processing task in terms of accuracy. An exception in the legal domain is represented by McCarty who in [16] developed a Definite Clause Grammar (DCG), consisting of approximately 700 rules, to result in "deep semantic interpretations" of a corpus of judicial opinions of the United States Supreme Court. He aimed at extracting the information that a lawyer wants to know about a case. He started from a qualitative analysis of general–purpose statistical parser (the Collins' parser) applied to those legal texts in order to test how accurate it was on sentences from judicial opinions. The parser results were mostly weak with respect to prepositional phrase attachments and coordinative conjunctions. Consequently, he foresaw several steps devoted to improving the accuracy of the parser for legal texts. It is also the case for Mochales and colleagues, who in [18] focussed on how legal language peculiarities are reflected in Argumentation Mining in legal texts. In order to detect and classify argumentative sentences in texts, the authors firstly looked for clauses of sentences on the basis of a predefined set of linguistic features, such e.g. typical verbs, significant sequences of two or more words, keywords, punctuation marks, etc. Then, using the linguistic characteristics of legal argument found in the previous phase, they defined a Context Free Grammar specifically devoted at parsing the argumentation structure in legal texts.

Finally, in the AI&Law field efforts devoted to investigating legal language peculiarities seem to be lacking. As will be demonstrated in the work presented here, such a study can help to shed light on those linguistic characteristics of legal documents, which might hamper Legal Knowledge Management efforts.

A significant exception is represented by the study by Nakamura and colleagues in [21], where the authors performed the linguistic investigation of a corpus of Japanese legal texts. Taking into account the linguistic characteristics they detected, they realised a system, which generates a logical formula corresponding to an input sentence from legal documents. In fact, they demonstrated that the results of this preliminary linguistic investigation are suitable for i) improving the accuracy of their NLP component that carries out the deep syntactic parsing of legal texts and ii) coping with particular legal sentence constructions that are difficult to transform into their corresponding logical forms. In particular, they put the focus on the analysis of typical Japanese nominalisations (i.e. noun phrases systematically related to the corresponding verbs), consisting of two nouns *A* and *B* with an adnominal particle "no", which carries some relation between *A* and *B*. The importance of an *'A no B'* relation relies on the fact that it is regarded as a verb. Frequently occurring in legal texts, these noun phrase types address the need for specific processing, in order to be transformed into a logical form, which expresses an event.

3 The Approach

This paper intends to continue the study carried out in Venturi [33] with a view to the practical importance that an investigation of the linguistic characteristics of legal texts has for Legal Knowledge Management purposes. In that previous study, the relative distribution of legal sub–language peculiarities has been identified by comparing the syntactic features detected in a corpus of Italian legal texts with the output of the syntactic parsing performed on a corpus of Italian ordinary language.

In the present study, a similar constrastive approach has been followed. Namely, syntactic and lexical characteristics of Italian and English legal language are identified by comparing an Italian and an English **legal** corpus with a reference corpus of Italian and English **ordinary** language. Afterwards, detected Italian and English legal language peculiarities are compared in order to investigate if, and to what extent, domain–specific characteristics are shared.

Syntactic and lexical levels of linguistic analysis have been carried out on the basis of the output of an NLP syntactic parsing component. In particular, the results presented here rely on a *shallow syntactic level of analysis*. As will be shown in Section 4, this paper maintains the widespread idea that a *shallow parsing* approach can help to provide enough detailed linguistic information for syntactically complex texts. Due to the minimal linguistic knowledge (i.e. morphosyntactic, lemma and word order information) a shallow syntactic component of analysis requires, such a level of analysis can be suitable to provide unambiguous syntactic representations.

4 NLP Analysis of Legal Texts

Syntactic and lexical levels of linguistic analysis are the focus of the present study. In particular, the latter level concerns *chunking*, the shallow syntactic

parsing technique, which segments sentences into an unstructured sequence of syntactically organised texts units called *chunks*. Abney in [1] demonstrated how chunking proves to be a highly versatile means to produce reliable syntactic annotations of texts. The purpose of traditional full–parsing is to associate to each sentence a fully specified recursive structure, in order to identify the proper syntagmatic composition, as well as the relations of functional dependency among the identified constituents. On the contrary, chunking refers to a process of non–recursive segmentation of text. The resulting analysis is flat and unambiguous: only those relations which can be identified with certainty have been found out. Accordingly, some of the ambiguous grammatical dependencies (e.g. prepositional phrase attachments) are left underspecified and unresolved. This makes chunking highly suitable for the syntactic annotation of different types of texts, both written and spoken, and the analysis of corrupted or fragmentary linguistic inputs. According to Li et al. [14], as long as "parse incompleteness" is reinterpreted as "parse underspecification", failures due to lexical gaps, particularly complex syntactic constructions, etc. are minimised.

A number of reasons for carrying out a *shallow parsing* of legal texts are the following. According to Li et al. [14], in many natural language applications, such as Information Extraction and Text Summarisation, it is sufficient to use shallow parsing information, rather then relying on a deep syntactic analysis.

Although it might seem that full parsing should be preferred for adequate processing of texts, a shallow parsing approach has been chosen within some domain–specific applications. This is the case, for example, for Grover and colleagues, who in [9] investigated a method for generating flexible summaries of legal documents, by detecting a set of argumentative roles (e.g. fact, background, proceedings, etc.). Relying on the output of a chunking component of analysis, the authors carried out a *fact extraction* task from a corpus of judgments of the House of Lords.

Moreover, Bartolini et al. [5] and Spinosa et al. [29] have shown in their works the main advantages in taking chunked syntactic structure as the basis on which further stages of legal text processing operate. It has been reported there that chunked representations can profitably be used as the starting point for partial functional analyses, aimed at reconstructing the range of dependency relations within the law paragraph text that are instrumental for the semantic annotation of text. The major potential for text chunking lies in the fact that chunking does not "balk" at the domain–specific constructions that do not follow general grammar rules; rather it actually carries on parsing, while leaving behind any chunk unspecified for its category.

5 Parsing Italian Legal Texts

5.1 The NLP Tools

AnIta (Bartolini et al. [4]) is the parsing system used for the analysis of Italian legal texts. It is a general–purpose parsing system, which has already been tested as a component both in the SALEM semantic annotation system of legal texts

(Bartolini et al. [5]) and in the MELT (Metadata Extraction from Legal Texts) system (Spinosa et al. [29]) showing encouraging results. *AnIta* is constituted by a pipeline of NLP tools, which also includes a chunking module, CHUG–IT (Federici et al. [7]). In CHUG–IT chunking is carried out through a finite state automaton which takes as input a morpho-syntactically tagged text. According to Federici et al. [7], a *chunk* is a textual unit of adjacent word tokens; accordingly, discontinuous chunks are not allowed. Word tokens internal to a chunk share the property of being mutually linked through dependency relations which can be identified unambiguously with no recourse to lexical information other than part of speech and lemma. A sample output of this syntactic processing stage is given in Figure 1, where the input sentence is segmented into four chunks. Please note that each chunk contains information about its type (e.g. a noun chunk, N_C, a finite verb chunk, FV_C, a prepositional chunk, P_C, etc.), its lexical head (identified by the label POTGOV) and any occurring modifier and preposition.

Le stesse disposizioni si applicano ad un prodotto importato
'The same provisions are applied to an imported product'
`[[CC:N_C][DET:LO#RD][PREMODIF:STESSO#A][POTGOV:DISPOSIZIONE#S]]`
`[[CC:FV_C][CLIT:SI#PQ][POTGOV:APPLICARE#V]]`
`[[CC:P_C][PREP:AD#E][DET:UN#RI][POTGOV:PRODOTTO#S]]`
`[[CC:ADJPART_C][POTGOV:IMPORTARE#V@IMPORTATO#A]]`

Fig. 1. CHUG–IT output

The chunked sentence in Figure 1, shows an example of the use of underspecification. The chunking process resorts to underspecified analyses in cases of systematic ambiguity, such as the one between adjective and past participle. This ambiguity is captured by means of the underspecified chunk category ADJPART_C, subsuming both an adjectival chunk and a participal chunk interpretation.

This underspecified approach to robust syntactic analysis of Italian texts has been proved to be fairly reliable. Lenci et al. [13] provides a detailed evaluation of CHUG–IT parsing performance drawn on a corpus of financial newspapers articles. Results of automatic chunking were evaluated against a version of the same texts chunked by hand; they give a recall of 90.65% and a precision of 91.62%.

In what follows we wil provide an analysis of a corpus of Italian legal texts. For this purpose, the output of the chunking module included in *AnIta* (i.e. CHUG–IT) has been analyzed.

5.2 The Corpora

For the construction of the Italian legislative corpora two different design criteria were taken into account, namely the regulated domain and the enacting authority. The corpus is made up of legal documents which a) regulate two different

domains, i.e. the environmental and the consumer protection domains and b) which are enacted by three different authorities, i.e. European Union, Italian state and Piedmont region.

The Environmental Corpus. The environmental corpus consists of 824 legislative, institutional and administrative acts for a total of 1,399,617 word tokens. It has been downloaded from the BGA (*Bollettino Giuridico Ambientale*), database edited by the Piedmont local authority for the environment[1]. The corpus includes acts enacted by the European Union, the Italian state and the Piedmont region, which cover a nine–year period (from 1997 to 2005). It is a heterogeneous document collection (henceforth referred to as Environmental Corpus) including legal acts such as national and regional laws, European directives, legislative decrees, as well as administrative acts, such as ministerial circulars and decision.

The Consumer Law Corpus. The corpus containing legal texts which regulate the consumer protection domain is a more homogeneous collection. Built and exploited in the DALOS project (Agnoloni et al. [2]), it is made up of 18 European Union Directives in consumer law (henceforth referred to as Consumer Law Corpus), for a total of 74,210 word tokens. Unlike the Environmental Corpus, it includes only Italian European legal texts.

5.3 Comparative Syntactic and Lexical Analysis

The investigation of syntactic and lexical peculiarities of legal language has been carried out starting from the chunked text (i.e. the output of CHUG–IT). The analysis mainly concerns:

1. the distribution of single chunk types;
2. the distribution of sequences of chunk types, with a view to those sequences which contain prepositional chunks;
3. the linguistic realization of events (i.e. situations) in legal texts.

A comparative method was followed. The distribution percentages of both single chunk types and sequences of chunks occurring within the Italian Legislative Corpus (i.e. the Environmental and the Consumer Law Corpus) were compared with the analysis of an Italian reference corpus, the PAROLE corpus (Marinelli et al., [15]), made up of about 3 million words including texts of different types (newspapers, books, etc.). Similarly, the typical linguistic realization of events in legal texts was highlighted by comparing the different lexical realization of situations depicted in legal documents and in the Italian reference corpus.

Distribution of Single Chunk Types. The distribution of single chunk types within legal texts was computed by comparing the occurrences of chunk types in the Italian Legislative Corpus and in the Italian reference corpus. This comparative approach is strengthened by the Chi-squared test applied on the obtained

[1] http://extranet.regione.piemonte.it/ambiente/bga/

results. It confirms the existence of a significant correlation between corpus variation and chunk type distribution.

Results of the parsing process, reported in Table 1, can help to highlight some main linguistic peculiarities of the Italian legal language and some consequences for Legal Knowledge Management. In particular, Table 1 shows the distribution of single chunk types in the Italian Legislative Corpus and in the Italian reference corpus. In this table, the count and the percentual frequency of occurrence are reported for each chunk type. It should be noted that the distribution of chunk types within the Environmental Corpus and the Consumer Law Corpus are kept distinct. As will be discussed in what follows, this choice of analysis brought about a number of related issues.

Table 1. Comparative distribution of single chunk types

Chunk types	Italian Legislative Corpus				PAROLE corpus	
	Environmental Corpus		Consumer Law Corpus			
	Count	%	Count	%	Count	%
Adj/Participial_C	38607	3.56	1689	2.74	29218	1.90
Adjectival_C	126267	11.66	6146	10.00	65740	4.27
Adverbial_C	13021	1.20	1006	1.63	49038	3.19
Coordinating_C	59585	5.50	3095	5.03	73073	4.75
Finite Verbal_C	36838	3.40	3007	4.89	140604	9.14
Nominal_C	226529	20.92	13062	21.25	413821	26.92
Non Finite Verbal_C	19569	1.80	5867	9.54	41674	2.71
Predicative_C	13047	1.20	843	1.37	21772	1.41
Prepositional_C	321167	29.66	14152	23.03	338037	21.99
Punctuation_C	192419	17.77	9756	15.87	278897	18.14
Subordinating_C	22026	2.03	2288	3.72	70226	4.56
Unknown_C	13439	1.24	535	0.87	14964	0.97

Interestingly enough, Table 1 shows that **prepositional chunks** (Prepositional_C) are the most frequent chunk types within the whole Italian Legislative Corpus. On the contrary, **nominal chunks** (Nominal_C) are the most recurring chunk types within the reference corpus. However, it should be appreciated that prepositional as well as nominal chunks are differently distributed between the Environmental Corpus and the Consumer Law Corpus. Namely, in the Environmental Corpus prepositional chunks constitute 29.66% of the considered chunks while the nominal chunks are 20.92%; in the Consumer Law Corpus the former ones are 23.03% while the latter ones are the 21.25%. Conversely, in the Italian reference corpus the nominal chunks are 26.92% of the total amount of chunk types and the prepositional chunks are 21.99%.

Moreover, a fairly low percentage of **finite verbal chunks** seems to be one of the main specific features of legal texts. Whereas the Italian reference corpus has 9.14% of the finite verbal chunks, their occurrence is about a third of that

within the Environmental Corpus, i.e. 3.40%, and they only constitute 4.89% of the total amount of considered chunk types in the Consumer Law Corpus.

Various remarks follow from the results obtained by this first level of shallow parsing. First, the different distributions of single chunk types within the two analysed corpora of legal texts raised the need for a finer–grained investigation of legal corpora. Such a further analysis took into account that this difference might be due to the different enacting authorities, i.e. the Italian state and the Piedmont region, which enacted two–thirds of the Environmental Corpus, and the European Union, which enacted both one–third of the Environmental Corpus and the whole Consumer Law Corpus. In order to investigate this hypothesis, we investigated the distribution of single chunk types within the three sub–corpora, which made the Environmental Corpus.

Table 2. Comparative distribution of single chunk types within three Environmental sub–corpora

Chunk Types	Italian Legislative Corpus					
	Environmental Corpus					
	Region		State		Europe	
	Count	%	Count	%	Count	%
Adj/Participial_C	7247	3.58	20305	3.58	11055	3.52
Adjectival_C	24949	12.33	68931	12.16	32387	10.33
Adverbial_C	2149	1.06	5944	1.04	4928	1.57
Coordinating_C	10315	5.09	31930	5.63	17340	5.53
Finite Verbal_C	5857	2.89	16601	2.92	14380	4.58
Nominal_C	42850	21.17	114404	20.18	69275	22.10
Non Finite Verbal_C	3509	1.73	7927	1.39	8133	2.59
Predicative_C	1850	0.91	6467	1.14	4730	1.50
Prepositional_C	59615	29.46	175011	30.87	86541	27.61
Punctuation_C	36373	17.97	103696	18.29	52350	16.70
Subordinating_C	3348	1.65	10068	1.77	8610	2.74
Unknown_C	4279	2.11	5496	0.96	3664	1.16

Results of this investigation are reported in Table 2, where the count and the percentual frequency of occurrence are shown for each chunk type. By keeping distinct the three different enacting authorities, different syntactic peculiarities of the legal language used in the European Italian legal texts and in the national and local legal texts were highlighted. Interestingly, it seems that the Italian European legal language has linguistic features which make it more similar to ordinary language than the national and local legal language. Table 2 shows in particular that the Environmental sub–corpus made up by legal texts enacted by the Italian state is characterised by the highest occurrence of prepositional chunks; these are 30.87% of the total amount of considered chunk types. They show a slightly lower occurrence in the Environmental sub–corpus made up by legal texts enacted by the Piedmont region, where the prepositional chunks are 29.46%, and it is 27.61% in the European part of the Environmental

Corpus. Moreover, the distribution of finite verbal chunks in the three sub–corpora strengthened the first hypothesis. They are 2.89% and 2.92% respectively in the local and in the national sub–corpus; while they occur twice as much in the European sub–corpus, i.e. 4.58%. Interestingly, this latter percentage distribution of finite verbal chunks is more similar to the corresponding distribution of this chunk type within the Italian reference corpus (i.e. 9.14%).

Thus, this comparative analysis resulted in a close relationship between the European Italian legal texts and the Italian reference corpus, closer than the relationship between the latter and the national and local legal documents. It seems to suggest that the European legislator, to a certain extent, took into account the frequently advocated plain language recommendations. In other words, the language used during the legal drafting process of European legal documents reveals itself as less different from ordinary language. It follows that the processing of European legal language may require fewer customizations of NLP tools due to legal language peculiarities than the processing of national and local legal texts. Consequently, Legal Knowledge Management applications in the European field will be less hampered by linguistic obstacles caused by domain–specific features.

Moreover, the two more visible syntactic peculiarities, i.e. the higher occurrence of prepositional chunks and the lower presence of finite verbal chunks, detected within the whole Italian Legislative Corpus with respect to the Italian reference corpus, raised the need for exploring two hypotheses. The first concerns the possibility that such a high occurrence of prepositional chunks is strongly connected with their presence within sequences of chunks. As it will be described in the "Distribution of Sequences of Chunk Types" Section, according to this hypothesis, the distribution of sequences of certain chunk types has been investigated. A special focus has been put on those sequences which contain prepositional chunks. The second hypothesis concerns the bias typical of legal texts towards a *nominal* realization of events (situations) rather than a *verbal* realization. The observed low occurrence of finite verbal chunks gave rise to this hypothesis. Accordingly, in the "Linguistic Realization of Events in Legal Texts" Section, an investigation will be carried out into how events are more typically expressed within the Italian Legislative Corpus with respect to the Italian reference corpus.

Distribution of Sequences of Chunk Types. The hypothesis made about by the high occurence of prepositional chunks within the Italian Legislative Corpus concerned the possibility that these chunk types would be typically contained in sequences of chunks. In particular, a hypothesis was put forward regarding the presence of long sequences which include a high number of embedded prepositional chunks.

In order to test this hypothesis, sequences of chunk types containing prepositional chunks have been automatically identified. The following typology of cases has been considered:

1. chains of consecutive prepositional chunks, such as the following excerpt *presentazione delle domande di contributo ai Comuni per l'attivazione dei*

distributori per la vendita di metano ([N_C presentazione] [P_C delle domande] [P_C di contributo] [P_C ai Comuni] [P_C per l'attivazione] [P_C di distributori] [P_C per la vendita] [P_C di metano]) "submission of contribution requests to Municipalities for the activation of distributors for the sale of natural gas";

2. sequences of prepositional chunks with possibly embedded adjectival chunks, such as the following excerpt *disciplina del canone regionale per l'uso di acqua pubblica* ([N_C disciplina] [P_C del canone] [ADJ_C regionale] [P_C per l'uso] [P_C di acqua] [ADJ_C pubblica]) "regulation of the regional fee for public water usage";
3. sequences of prepositional chunks with possibly embedded adjectival chunks, coordinative conjunctions and/or "light" punctuation marks (i.e. comma), such as the following excerpt *acqua destinata all'uso igienico e potabile, all'innaffiamento degli orti ...* ([N_C acqua] [ADJPART_C destinata] [P_C all'uso] [ADJ_C igienico] [COORD_C e] [ADJ_C potabile] [PUNC_C,] [P_C all'innaffiamento] [P_C degli orti]) "water devoted to sanitary and drinkable usage, to garden watering".

The investigation especially focused on the different distribution of deep chains containing prepositional chunks (referred to as *PP–chains*) in the different kinds of texts considered. Results are shown in Table 3, which shows the count of embedded PP–attachments (i.e. sequences of chunk types containing embedded prepositional chunks) that occurred within a sentence of legal texts with respect to an ordinary language sentence [2].

By inspecting Table 3, the occurrence of deep PP–chains does not prove to be a special syntactic feature of legal language with respect to ordinary language. Rather, the crucial distinguishing characteristic of the Italian Legislative Corpus appears to be the *different percentual distributions of deeply embedding sequences containing prepositional chunks*. Legal texts appear to have a higher percentage of deep PP–chains with respect to the Italian reference corpus. Moreover, the analysis of different percentual occurrences within the three different Environmental sub–corpora and within the Consumer Law Corpus allowed the highlighting of finer–grained peculiarities. In general, it should be noticed that there mainly are chains including 5 to 11 embedded chunks. For example, chains of 8 PP–attachments constitute 5.78% of the total amount of PP-chains occurring within the legal texts enacted by the Piedmont region and 5.52% in the documents enacted by the Italian state. Yet, theyhave a coverage of only 2.47% in the Italian reference corpus. As highlighted in the "Distribution of Single Chunk Types" Section, the Italian European legal texts show a close relationship with ordinary language. Accordingly, chains of 8 PP–attachments have lower frequency of occurrence; they are 4.24% in the European part of the Environmental Corpus and 3.26% in the Consumer Law Corpus.

These findings allow us to consider a number of statements. First, deep PP–attachment chains seem to be typical of legal texts. They range from chains of

[2] Note that the first column of the Table above (named "PP–chains depth") reports the number of chunk types embedded, with respect to the typology of cases considered.

Table 3. Comparative distribution of PP–attachment chains

PP–chains depth	Italian Legislative Corpus								PAROLE Corpus	
	Environmental Corpus						Consumer Law Corpus			
	Region		State		Europe					
	Count	%	Count	%	Count	%	Count	%	Count	%
4	2822	38.48	8924	37.42	4164	43.19	611	45.32	10240	54.72
5	1723	23.71	5366	22.50	2258	23.42	356	26.40	4621	24.68
6	1043	14.35	3505	14.69	1380	14.31	139	10.31	1999	10.68
7	612	8.42	2103	8.81	725	7.52	104	7.75	910	4.85
8	420	5.78	1318	5.57	409	4.24	44	3.26	464	2.47
9	248	3.41	813	3.40	237	2.45	28	2.07	206	1.09
10	151	2.13	652	2.73	161	1.67	23	1.70	112	0.59
11	91	1.35	350	1.46	92	0.95	10	0.74	74	0.39
12	63	0.88	244	1.02	69	0.71	7	0.51	39	0.20
13	30	0.42	167	0.70	39	0.40	9	0.66	28	0.14
14	19	0.32	147	0.61	37	0.38	5	0.37	17	0.09
15	18	0.28	79	0.33	27	0.28	1	0.07	6	0.03
16	11	0.25	62	0.25	26	0.27	6	0.44	5	0.02
17	6	0.09	40	0.16	5	0.05	1	0.07	3	0.01
18	3	0.05	31	0.12	4	0.04	3	0.22	2	0.01
19	3	0.04	24	0.10	3	0.03	0	0.00	1	0.00
20	2	0.02	23	0.09	4	0.04	1	0.07	3	0.00

embedded cross–reference to other legal documents, or sections of text (such as paragraphs, articles, etc.), such as the following sequence containing embedded prepositional chunks *all'articolo 1, comma 1, della legge 8 febbraio 2001, n. 12, ...* "in article 1, paragraph 1, of the act 8 February 2001, n. 12, ...", to chains of deverbal nouns (i.e. nouns morphologically derived from verbs), such as the following example *ai fini dell'accertamento della compatibilità paesaggistica ...* "to the verification of the landscape compatibility ...". In both cases, detecting these kinds of deep PP–chains would be fruitful for legal document transparency. As a matter of fact, the recurrence of complex and ambiguous syntactic constructions, such as deep sequences of prepositional chunks, is widely acknowledged to be responsible for the lack of understandability of legal texts. According to Mortara Garavelli [20], it is not the occurrence of abstract deverbal nouns which may affect the whole legal text comprehension; rather, the complex syntactic patterns, in which these deveba nouns are typically embedded, make legal texts difficult to comprehend. This is in line with some findings in studies on linguistic complexity, mainly in the cognitive and psycholinguistic field (see Fiorentino [8] for a survey of the state–of–the–art). It was discovered that our short term memory is able to receive, process and remember an average of 7 linguistic units. In processing a given input sentence the language user attempts to obtain closure on the linguistic units contained in it as early as possible. Thus, it is perceptually "costly" to carry on analysing deep chains of embedded sentence constituents.

Finally, as mentioned above, the analysis of sequences of prepositional chunks containing **deverbal nouns** may be related to a study of the linguistic realization of events (situations) in legal texts. Let us consider the two following sentences:

1. l'autorità amministrativa competente accerta la compatibilità paesaggistica ("the relevant administrative authority verifies the landscape compatibility"),
2. il Comune è preposto alla gestione del vincolo ai fini dell'accertamento della compatibilità paesaggistica ... ("the Municipality is in charge of the management of the obligation to the verification of the landscape compatibility").

In the first case, the event "verification" is realised through a verbal construction involving the verb *accertare* ('to verify'). In the second sentence, the same event is realised through a nominal construction headed by the deverbal noun *accertamento* ('verification'). Interestingly, it should be noted that in the latter case the deverbal noun is embedded in a sequence of prepositional chunks, i.e. *... preposto alla gestione del vincolo ai fini dell'accertamento della compatibilità paesaggistica ...* "... in charge of the management of the obligation to the verification of the landscape compatibility ...". According to these findings, remarks on Legal Knowledge Management applications such as Event Extraction can benefit from the results obtained by an analysis of PP–chains containing deverbal nouns.

Linguistic Realization of Events in Legal Texts. The low percentual occurrence of finitive verbal chunks found in Section 5.3 hinted at lexical realization patterns of situations and events, which is typical of legal documents. In order to follow this direction of research, a case study was carried out on a small sample of some main events within the Italian Legislative Corpus and the Italian reference corpus.

The results are reported in Table 4, where for each event type the corresponding verbal and nominal morpho–syntactic realization is shown in the second column. It should be noted that the percentual occurrence of the type of morpho–syntactic realization has been computed as the ratio of the noun (or of the verb) occurrence over all types of realization (i.e. nominal + verbal) of a given event. In the last columns of the table, the count and the percentual occurrence of the two linguistic realization types are shown. Interestingly, it highlights a broad bias towards a **nominal** realization of same main events within the Italian Legislative Corpus.

This is the case for the 'Violate' event triggered by words which convey a situation where someone or something violates a set of rules. As shown in Table 4, this event type can be expressed by the verb 'violare' (to violate) and by the deverbal noun 'violazione' (infringement). **Nominal** realization was more frequent in the Italian Legislative Corpus than in the Italian reference corpus. However, a different percentual occurrence can be seen in the legal texts enacted by the European Union (both regulating the environmental and consumer protection domain), and in the documents enacted by the local authority and by the Italian state. According to previous findings, the local and national legal

Table 4. Comparative morpho–syntactic realization of events

Event type	Morpho-syntactic realization	Italian Legislative Corpus				PAROLE Corpus	
		European texts		Regional & national texts			
		Count	%	Count	%	Count	%
ENFORCE	attuare (to enforce)	159	24.02	184	9.94	88	43.35
	attuazione (enforcement)	503	75.98	1668	90.06	115	56.65
VIOLATE	violare (to violate)	8	9.09	5	2.94	107	52.97
	violazione (infringement)	80	90.91	165	97.06	95	47.03
PROTECT	proteggere (to protect)	107	16.61	296	26.35	179	55.59
	protezione (protection)	537	83.39	819	73.45	143	44.41
IMPOSING_OBLIGATION	obbligare (to obligate)	19	6.01	59	8.18	122	42.21
	obbligo (obligation)	297	93.99	662	91.82	167	57.79

texts seem to contain more domain–specific peculiarities. In those documents, the 'Violate' event is realized in 97.06% of the total amount of cases by the deverbal noun 'violazione' (infringement) and only in 2.94% of cases by the verb 'violare' (to violate). In the European documents there is also a strong bias towards the nominal realization, however with different occurrence percentages: the deverbal noun 'violazione' (infringement) occurs in 90.91% of all 'Violate' event realizations, and the verb 'violare' (to violate) is 9.09% of cases.

Conversely, within the Italian reference corpus the variance of morpho–syntactic realization type shows different characteristics. Not only is the verbal realization more frequent than the nominal one – 52.97% versus 47.03% respectively –, but also, it seems that ordinary language does not have any sharp bias towards one of the two types of linguistic realization.

These findings prompted an assessment of the consequences for Legal Knowledge Management tasks such as Event Extraction from legal document collections. According to the state–of-the–art literature in the Event Knowledge Management field, the processing of nouns and deverbal nouns is as crucial as challenging. As it is claimed in Gurevich et al. [10], deverbal nouns, or *nominalizations*, pose serious challenges for general–purpose knowledge–representation systems. They report that the most common strategy to face with this relevant problem involves finding ways to create verb–like representations from sentences which contain deverbal nouns, i.e. strategies to map the arguments of deverbal nouns to those of the corresponding verbs. In particular, tasks such as Semantic Role Labelling for event nominalizations [22] are very concerned with this challenge.

The results shown in this section reveal that, more than within the open–domain field, the Event Knowledge Management task in the legal domain is made more challenging by the rather high occurrence of nouns and deverbal nouns, which should be considered *event predicative*.

6 Parsing English Legal Texts

The syntactic and lexical analysis of English Legal texts adopted the same criteria applied to the investigation of Italian legal texts. Accordingly, it relies on a shallow level of syntactic parsing (i.e. chunking), and it is carried out by following a similar comparative method. As in the case of the linguistic analysis of Italian legal language, the relative distribution of English legal language characteristics has been investigated with respect to an English reference corpus. As it will be described in Section 7, comparing the results obtained by parsing Italian and English legal texts, the eventual goal is to investigate whether some syntactic and lexical peculiarities were shared by the Italian and English legal language.

6.1 The NLP Tools

The *GENIATagger* (Tsuruoka et al. [30]), a NLP component carrying out part–of–speech tagging and chunking, has been used to perform the English legal texts analysis. Even though the output of this component is quite similar to CHUG–IT's, the output of the two tools differs to some extent. In fact, they mainly diverge because of different grammatical requirements in the two languages considered (i.e. Italian and English), as well as differences in linguistic annotation choices.

The fragment of *GENIATagger* chunked text, reported in Table 5, shows how the *GENIATagger* outputs. In the first column (Word Form) the word is shown as it appears in the original sentence; the second column lists the lemma of the word; the part–of–speech tag is in the third column (e.g. NN is the tag used for nouns, IN is the tag which labels prepositions other than *to*, etc.). The last column indicates the chunk type (e.g. NP indicated a nominal chunk, PP is a prepositional chunk, etc.). It should be noted that chunks are represented in the IOB2 format; thus, in the Table B stands for BEGIN (of the chunk) and I for INSIDE (the chunk itself).

In particular, it should be noticed that the output of the *GENIATagger* and that of CHUG–IT mostly differ because of their representation of nominal and prepositional chunks. A prepositional chunk does not contain anything more than a preposition inside, such as *to* or *as* which are at the beginning of the PP chunk (i.e. B–PP). This annotation strategy is relevant for the English syntactic features concerning the stranding of prepositions within a sentence. Conversely, a nominal chunk can be a textual unit of adjacent word tokens, such as *certain exonerating circumstances*, which includes an adjective (*certain*, JJ) at the beginning (B–NP), an introducing present participle (*exonerating*, VBG) and a common noun (*circumstances*, NNS) as two inner elements (I–NP). Yet, it can

also be made up of a single word token, such as *proof*, which includes a common noun (NN) only, or *he*, which is made up of a personal pronoun (PRP). However, it can never include a preposition.

On the contrary, the annotation strategy of CHUG–IT allows segmenting a sentence differently. As reported in Section 5.1, for example, the prepositional chunk *ad un prodotto*, "to a product", always includes both the preposition *a* ("to") and the noun *prodotto* ("product").

Table 5. GENIATagger annotation

Word Form	Lemma	Part-Of-Speech	Chunk Type
he	he	PRP	B-NP
proof	proof	NN	B-NP
furnishes	furnishes	VBZ	B-VP
as	as	IN	B-PP
to	to	TO	B-PP
the	the	DT	B-NP
existence	existence	NN	I-NP
of	of	IN	B-PP
certain	certain	JJ	B-NP
exonerating	exonerating	VBG	I-NP
circumstances	circumstances	NNS	I-NP

"... he furnishes proof as to the existence of certain exonerating circumstances ..."

6.2 The Corpus

For the English legal text analysis, a collection of 18 English European Union Directives in consumer law has been used. The corpus has been built and exploited in the DALOS project (Agnoloni et al. [2]). It is made up of the English version of the Italian corpus in consumer law. This legal corpus has been compared with a sub–corpus of the Wall Street Journal made up of 1,138,189 words, which was used as a reference corpus.

6.3 The Comparative Syntactic and Lexical Analysis

Differently from the Italian case, the comparison between the English Legislative Corpus and the reference corpus (i.e. WSJ Corpus) has concentrated on:

1. the distribution of single chunk types,
2. the linguistic realization of events (i.e. situations) in legal texts.

A more exhaustive syntactic investigation is still ongoing, also including the analysis which concerns the distribution of sequences of chunk types, compared to those sequences which contain prepositional chunks.

Distribution of Single Chunk Types. The distribution comparison of chunk types, between the English Legislative Corpus in consumer law and the

reference corpus, shows some legal language peculiarities wich have been detected previously for Italian. As in the Italian legal texts, within the English legal documents the occurrence of **prepositional chunks** has been noted to be higher than in the general language texts (see Table 6). They constitute 27.21% of the total number of chunk types in the English European Union Directives, against 19.88% in the Wall Street Journal sub–corpus. At the same time, the percentage of **nominal chunks** is lower in legal texts (48.16%) than in the reference corpus, where they represent 51.84% of the identified chunks.

Regarding the distribution of **finite verbal chunks**, the comparative analysis shows that they have a quite low percentage of occurrence. In particular, they represent 9.17% of the total chunk types within the English legislative corpus, compared to 15.56% in the reference corpus.

Table 6. Comparative distribution of chunk types

Chunk Types	English Legislative Corpus		WSJ Corpus	
	Count	%	Count	%
Nominal_C	17731	48.16	336635	51.84
Prepositional_C	10019	27.21	129131	19.88
Finite verbal_C	3378	9.17	101092	15.56
Non finite verbal_C	2401	6.52	26673	4.10
Adverbial_C	835	2.26	24139	3.71
Adjectival_C	823	2.23	11726	1.80

It should be noted that these results are in line with the ones from the analysis of the corpus of Italian legal texts enacted by the European Community. In fact, one of the most prominent findings in Section 5.3 was the close relationship between European Italian legal language and ordinary language. In that case, it was shown that such relationship is closer than the one between the legal language used in national and local documents, and ordinary Italian. However, the rather low frequency of finite verbal chunks found in European English legal texts suggest that these documents are possibly characterised by a significant bias towards a nominal realization of events.

Linguistic Realization of Events in Legal Texts. In order to investigate the hypothesis motivated by the low occurrence of finite verbal chunks in the English Legislative Corpus, a case study was carried out on a small sample of some main events. Similar to the Italian case study, the different lexical realizations of situations depicted in English legal texts and in the English reference corpus were investigated. The percentual occurrence of each type of morpho–syntactic realization was computed as the ratio of the noun (or of the verb) occurrence over all types of realization (i.e. nominal + verbal) of a given event. The results reported in Table 7 verify the first hypothesis, i.e. a broad bias typical of English legal documents towards the **nominal** realization of events.

In fact, in most of the cases reported in Table 7, the event nominal realizations are percentually more frequent in the legislative corpus than the verbal

constructions; while an opposite bias has been observed within the WSJ reference corpus. Interestingly, as previously observed for the Italian case, the same 'Violate' event within the English Legislative Corpus is realized in 2.67% of the total amount of occurrences through the verb 'to violate', while the nominal realization through the noun 'infringement' is 97.33% of cases. On the contrary, within the WSJ Corpus the event occurs more frequently through a verbal construction, i.e. 86.59% of cases, than through a nominal one, i.e. 13.41%.

According to these findings, as has previously been discussed in see Section 5.3, the investigation of the linguistic realization of events in legal texts might be of great importance for Event Knowledge Management.

Table 7. Comparative morpho–syntactic realization of events

Event type	Morpho-syntactic realization	English Legislative Corpus		WSJ Corpus	
		Count	%	Count	%
ENFORCE	to enforce	8	14.81	17	26.56
	enforcement	46	85.19	47	73.44
VIOLATE	to violate	2	2.67	71	86.59
	infringement	73	97.33	11	13.41
PROTECT	to protect	64	27.35	116	45.14
	protection	170	72.65	141	54.86
PROHIBIT	to prohibit	10	23.26	40	80.00
	prohibition	33	76.74	10	20.00

7 Comparing Italian and English Legal Language Peculiarities

In order to investigate which syntactic and lexical peculiarities are shared by the Italian and English legal language, we compare the results obtained in Section 5 and Section 6 with respect to:

1. the distribution of single chunk types,
2. the linguistic realization of events (i.e. situations) in legal texts.

This multilingual comparison takes into account the results obtained by contrasting the sub–part of the Environmental Corpus made up of texts enacted by the European Union and the Italian Consumer Law Corpus with the English Legislative Corpus. This was made possible by the homogeneous nature of the three corpora: they are all enacted by the same enacting authority, i.e. the European Union. It follows that this comparison concerns those syntactic and lexical peculiarities shared by the European legal language used in the two considered corpora.

Despite the different grammatical requirements in the two languages considered (i.e. Italian and English) and the different annotation choices of the two NLP tools exploited (i.e. the *GENIATagger* and CHUG–IT), the distribution

of single chunk types between Italian and English European legal texts shows similarities. Comparing the two European legal languages, some main features have been revealed as shared, namely:

1. a high occurrence of prepositional chunks,
2. a fairly low presence of finite verbal chunks.

The percentual distribution of these two chunk types within the considered legal corpora with respect to the corresponding distribution within the analysed reference corpora is significantly similar in the Italian and English cases. Namely:

1. within the European sub–part of the Italian Environmental Corpus, the prepositional chunks represent 27.61% of the total amount of chunks and 23.03% in the Italian Consumer Law Corpus; on the contrary, this chunk type coverss 21.99% in the Italian reference corpus;
2. within the European sub–part of the Italian Environmental Corpus, the finite verbal chunks cover 4.58% of the total amount of chunks and 4.89% in the Italian Consumer Law Corpus; on the contrary, this chunk type counts for 9.14% in the Italian reference corpus;
3. within the English Legislative Corpus, the prepositional chunks represent 27.21% of the total amount of chunks computed; on the contrary, this chunk type constitutes 19.88% in the English reference corpus;
4. within the English Legislative Corpus, the finite verbal chunks cover 9.17% of the total amount of chunks; on the contrary, this chunk type covers 15.56% in the English reference corpus.

Interestingly enough, it has been shown that both in the Italian case (see Section 5.3) and in the English case (see Section 6.3) the fairly low presence of finite verbal chunks within legal texts is closely related to a typical linguistic realization of events. However roughly, this shallow level of syntactic analysis shows a shared broad bias towards a *nominal* realization of some main events within Italian and English European legal texts.

8 Conclusion and Future Directions of Research

The results of an analysis of the main syntactic features of legal language detected within legal corpora have been presented in this article. Such an investigation relies on the output of an NLP component of analysis, which syntactically parses document collections at a *shallow level*. This output of *chunking components* has been analysed, and a three–level comparison has ben performed:

1. specialised and ordinary language,
2. different legal languages used by different enacting authorities (i.e. European Union, Italian state and Piedmont region),
3. two different European legal languages (i.e. Italian and English), assuming a multilingual perspective.

Even if quite rudimentary, this first level of syntactic grouping has allowed us to detect some main characteristics of legal language.

Among others, the quite high occurrence of prepositional chunks and the fairly low presence of finite verbal chunks have been considered as two of the more visible syntactic phenomena particular to legal sub–language. Interestingly, these main syntactic peculiarities observed are shared by the two languages considered (i.e. Italian and English).

The investigation of the reason why finite verbal chunks occur with a low frequency within legal texts led us to detect a broad bias within both Italian and English legal corpora towards a **nominal** realization of events rather than verbal. In the article, it has been pointed out how the outcome of such a linguistic study can have practical importance for Legal Knowledge Management tasks, such as Event Knowledge Management. According to the state–of–the–art literature, the nominal realization of events poses serious challenges for knowledge–representation systems. Consequently, the rather high occurrence of nominal realizations within legal corpora might cause difficulties for Event Knowledge Management in the legal domain. This is in line with the work carried out by Nakamura and colleagues in [21]. They started from the results obtained during a phase of investigation of the linguistic realization of events within Japanese legal texts. In particular, they put the focus on the analysis of typical Japanese nominalisations (i.e. noun phrases systematically related to the corresponding verbs) that frequently occur within legal texts. Consequently, the authors relied on a specific processing of these noun phrase types in order to transform them into a logical form which expresses an event.

According to the results described in the "Distribution of Sequences of Chunk Types" Section, the high occurrence of prepositional chunks found in the Italian Legislative Corpus appears to be related to the bias towards nominal realization of events within legal texts. In particular, according to the typology of chains of prepositional chunks, deep sequences of prepositional chunks containing deverbal nouns were found to be connected with the nominal realization of events. It has been pointed out how these findings can have practical importance for Event Knowledge Management.

Moreover, the high frequency of deep sequences of prepositional chunks has been related with the lack of understability of legal texts. These outcomes suggest how legal texts are difficult to be understood not only by human beings, but also by computational tools. In particular, deep PP–attachment chains can pose serious challenges for an NLP syntactic component in charge of a *dependency level* of syntactic analysis of Italian legal texts. In particular, the different syntactic aspects of legal corpora analysed when compared to the Italian reference corpus dramatically suggests the need for a training phase of NLP tools for deep parsing purposes.

Acknowledgements

This work is the result of research activities of the Dylan Lab (Dynamics of Language Laboratory) of the Institute of Computational Lnguistics (ILC–CNR)

of Pisa and of the Department of Linguistics (Computational Linguistics division) of the University of Pisa. *AnIta* has been developed and provided by its members.

References

1. Abney, S.P.: Parsing by chunks. In: Berwick, R.C., et al. (eds.) Principled–Based Parsing: Computation and Psycholinguistics, pp. 257–278. Kluwer Academic Publishers, Dordrecht (1991)
2. Agnoloni, T., Bacci, L., Francesconi, E., Peters, W., Montemagni, S., Venturi, G.: A two–level knowledge approach to support multilingual legislative drafting. In: Breuker, J., Casanovas, P., Francesconi, E., Klein, M. (eds.) Law, Ontologies and the Semantic Web. IOS Press, Amsterdam (2009)
3. Ananiadou, S., McNaught, J. (eds.): Text Mining for Biology and Biomedicine. Artech House, London (2006)
4. Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V.: Hybrid Constraints for Robust Parsing: First Experiments and Evaluation. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation, Centro Cultural de Belem, Lisbon, Portugal, May 26-28, pp. 795–798 (2004)
5. Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V., Soria, C.: Automatic classification and analysis of provisions in legal texts: a case study. In: Meersman, R., Tari, Z., Corsaro, A. (eds.) OTM-WS 2004. LNCS, vol. 3292, pp. 593–604. Springer, Heidelberg (2004)
6. Brighi, R., Lesmo, L., Mazzei, A., Palmirani, M., Radicioni, D.P.: Towards Semantic Interpretation of Legal Modifications through Deep Syntactic Analysis. In: Proceedings of the 21th International Conference on Legal Knowledge and Information Systems (JURIX 2008), Florence, December 10-13 (2008)
7. Federici, S., Montemagni, S., Pirrelli, V.: Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In: Carroll, J. (ed.) Proceedings of the Workshop on Robust Parsing, ESSLLI, Prague, August 12-16 (1996)
8. Fiorentino, G.: Web usability e semplificazione linguistica. In: Venier, F. (ed.) Rete Pubblica. Il dialogo tra Pubblica Amministrazione e cittadino: linguaggi e architettura dellinformazione, Perugia, Edizione Guerra, pp. 11–38 (2007)
9. Grover, C., Hachey, B., Hughson, I., Korycinski, C.: Automatic Summarisation of Legal Documents. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003), Scotland, United Kingdom, pp. 243–251 (2003)
10. Gurevich, O., Crouch, R.: Deverbal Nouns in Knowledge Representation. Journal of Logic and Computation 18(3), 385–404 (2008)
11. Lame, G.: Using NLP techniques to identify legal ontology components: concepts and relations. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 169–184. Springer, Heidelberg (2005)
12. Lease, M., Charniak, E.: Parsing Biomedical Literature. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 58–69. Springer, Heidelberg (2005)
13. Lenci, A., Montemagni, S., Pirrelli, V.: CHUNK–IT. An Italian Shallow Parser for Robust Syntactic Annotation. In: Linguistica Computazionale, Istituti Editoriali e Poligrafici Internazionali, Pisa–Roma (2001)

14. Li, X., Roth, D.: Exploring Evidence for Shallow Parsing. In: Proceedings of the Annual Conference on Computational Natural Language Learning, Toulouse, France (2001)
15. Marinelli, R., Biagini, L., Bindi, R., Goggi, S., Monachini, M., Orsolini, P., Picchi, E., Rossi, S., Calzolari, N., Zampolli, A.: The Italian PAROLE corpus: an overview. In: Zampolli, A., Calzolari, N., Cignoni, L. (eds.) Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVI-XVII, Pisa–Roma, IEPI. Tomo I, pp. 401–421 (2003)
16. McCarty, L.T.: Deep Semantic Interpretations of Legal Texts. In: Proceedings of the 11th international conference on Artificial intelligence and law (ICAIL), June 4-8. Stanford Law School, Stanford (2007)
17. McCarty, T.: Remarks on Legal Text Processing – Parsing, Semantics and Information Extraction. In: Proceedings of the Worskhop on Natural Language Engineering of Legal Argumentation, Barcelona, Spain, June 11 (2009)
18. Mochales Palau, R., Moens, M.-F.: Argumentation Mining: the Detection, Classification and Structure of Arguments in Text. In: Proceedings of the 12th international conference on Artificial intelligence and law (ICAIL), June 8-12. Universitat Autonoma de Barcelona, Barcelona (2009)
19. Moens, M.-F., Mochales Palau, R., Boiy, E., Reed, C.: Automatic detection of Arguments in Legal Texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and law (ICAIL), June 4-8. Stanford Law School, Stanford (2007)
20. Mortara Garavelli, B.: Le parole e la giustizia. In: Divagazioni grammaticali e retoriche su testi giuridici italiani, Torino, Einaudi (2001)
21. Nakamura, M., Nobuoka, S., Shimazu, A.: Towards Translation of Legal Sentences into Logical Forms. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) JSAI 2007. LNCS (LNAI), vol. 4914, pp. 349–362. Springer, Heidelberg (2008)
22. Pado, S., Pennacchiotti, M., Sporleder, C.: Semantic role assignment for event nominalisations by leveraging verbal data. In: Proceedings of Coling 2008, Manchester, UK, August 18-22 (2008)
23. Pala, K., Rychlý, P., Šmerk, P.: Automatic Identification of Legal Terms in Czech Legal Texts. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 83–94. Springer, Heidelberg (2010)
24. Pala, K., Rychlý, P., Šmerk, P.: Morphological Analysis of Law texts. In: Sojka, P.H., Aleš-Sojka, P. (eds.) First Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2007), pp. 21–26. Masaryk University, Brno (2007)
25. Pyysalo, S., Salakoski, T., Aubin, S., Nazarenko, A.: Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics 7(Suppl. 3), S2 (2006)
26. Rissland, E.L.: Ai and legal reasoning. In: Proceedings of the International Joint Conference in Artificial Intelligence, IJCAI 1985 (1985)
27. Sagae, K., Tsujii, J.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Proceedings of EMNLP–CoNLL 2007, pp. 1044–1050 (2007)
28. Saias, J., Quaresma, P.: A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. In: Benjamins, V.R., Casanovas, P., Breuker, J., Gangemi, A. (eds.) Law and the Semantic Web. LNCS (LNAI), vol. 3369, pp. 185–200. Springer, Heidelberg (2005)

29. Spinosa, P., Giardiello, G., Cherubini, M., Marchi, S., Venturi, G., Montemagni, S.: NLP–based Metadata Extraction for Legal Text Consolidation. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009), Barcelona, Spain, June 8-12 (2009)
30. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part–of–Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 382–392. Springer, Heidelberg (2005)
31. Völker, J., Langa, S.F., Sure, Y.: Supporting the Construction of Spanish Legal Ontologies with Text2Onto. In: Casanovas, P., Sartor, G., Casellas, N., Rubino, R. (eds.) Computable Models of the Law. LNCS (LNAI), vol. 4884, pp. 105–112. Springer, Heidelberg (2008)
32. Van Gog, R., Van Engers, T.M.: Modelling Legislation Using Natural Language Processing. In: Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2001 (2001)
33. Venturi, G.: Linguistic analysis of a corpus of Italian legal texts. A NLP–based approach. In: Karasimos, et al. (eds.) Proceedings of First Patras International Conference of Graduate Students in Linguistics (PICGL 2008), pp. 139–149. University of Patras Press (2008)
34. Walter, S., Pinkal, M.: Automatic extraction of definitions from german court decisions. In: Proceedings of the COLING 2006 Workshop on Information Extraction Beyond The Document, Sydney, July 2006, pp. 20–28 (2006)
35. Walter, S.: Linguistic Description and Automatic Extraction of Definitions from German Court Decisions. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May 28-30 (2008)
36. Wyner, A., Mochales-Palau, R., Moens, M.F., Milward, D.: Approaches to Text Mining Arguments from Legal Cases. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 60–79. Springer, Heidelberg (2010)
37. Wyner, A., van Engers, T.: From Argument in Natural Language to Formalised Argumentation: Components, Prospects and Problems. In: Proceedings of the Worskhop on Natural Language Engineering of Legal Argumentation, Barcelona, Spain, June 11 (2009)